Lancaster Site Report

HEPSYSMAN @ RAL
May 11th 2012

Matt Doidge, on behalf of Lancaster Team.

The Lancaster HEP Squad

Starring:
Alex Finch
Matt Doidge
Peter Love

and introducing
Robin Long

The Local Cluster

- Usual services: (mainly) SL5 desktops, nfs home areas, tape backups, web & mail.
- Gateway to the Grid: A few UI boxes (which users keep doing development work on) and local access to cvmfs. Atlas & t2k users have access to large nfs "scratch" areas.
- 100% PROOF: Peter runs a PROOF cluster for the local. Four Dell C6100s delivering 96 cores (x5650s). The local users heavily use this service.

The Grid Side.

A tale of two clusters: 512 cores in the "HEP" and a ~1600 core share of the University "HEC" cluster.

- Both clusters host multiple generations of kit.
- Both are behind CREAM CEs.
- HEP uses torque/maui, HEC uses LSF for the batch system.

Shacking Up: We've moved almost all of our kit to the same machine room, out of our old, dank basement.

Grid Storage

All of our ~Petabyte of storage is in 24-bay supermicros with 1 or 2-TB disks. Mix of areca, 3ware and adaptec raid cards - monitoring is a dog.

- Procurements split between HEC & HEP subnets.

Our storage is behind a (gLite) DPM headnode on solid hardware.

- Ran well, before the recent DPM troubles. Pending a bug fix upgrade.

That Network Upgrade...

The mad scramble network uplift plan for Lancaster took a 3-pronged approach.

1. Upgrade & Shang-hai the University's back up link. 10G (mostly) just for us.
2. Increase connectivity to campus backbone & thus between the two "halves" of the grid cluster and the local HEP cluster.
3. Add capacity for 10G networking to our cluster using a pair of Z9000 core switches & half a dozen S4810 rack switches.
    a. This free's up some of the current switches that can be retasked to improve the HEP cluster networking.

The Whole of the Node.

Our most recent adventure was enabling whole node jobs on our cluster.

- Assigned a portion of our (lower memory) worker nodes exclusively for multicore work.
- Seemed pretty straightforward to set up in torque, although maui was a little off till I partitioned off the single core nodes & multi-core job nodes.

The Rest of the Whole of the Node.

- Andy at Edinburgh is having a go at testing our cluster.
- No news yet if things are working as intended (or at all).
- But seemed straight forward to set up, a good way breathing extra life into older, lower memory nodes.

The Watchers in the Dark.

We're currently undergoing a review of our (local) monitoring.

- Currently using ganglia, nagios, syslogs & hardware alerts.
  - We also have an underused OSSEC instance.
- Plan to pull more tests into a new nagios instance, located outside our domain.
- Keen to move to a "Measure Anything, Measure Everything" approach: Use statd to collect data from all & sundry and publish it using Graphite.

Deployment Tactics.

- Used to use cfengine to configure site
  - then our cfengine host died.
- We started living without cfengine, and found it not too bad.
  - Kickstart & a handful of setup scripts.
- Debating whether or not to give puppet a try (all the cool kids are doing it).
- For our shared cluster we're stuck using kusu, which is a pain but trying to get two configuration managers to play nice seems like an exercise in futility (if not outright masochism).

Questions!

- We have a number of services on virtual machines, but our VM server is poorly - anyone have any experience migrating KVM images?
- We've had a habit of using any old boxen to serve up our WN tarballs & SW area. Has anyone seen an advantage in using dedicated, purpose specced HW for this task?
- A possible third question.....